



Europäisches
Patentamt

European
Patent Office

Office européen
des brevets

PCT/IB03/05601

Bescheinigung

Certificate

Attestation

REC'D 24 DEC 2003

WIPO

PCT

Die angehefteten Unterlagen stimmen mit der ursprünglich eingereichten Fassung der auf dem nächsten Blatt bezeichneten europäischen Patentanmeldung überein.

The attached documents are exact copies of the European patent application described on the following page, as originally filed.

Les documents fixés à cette attestation sont conformes à la version initialement déposée de la demande de brevet européen spécifiée à la page suivante.

Patentanmeldung Nr. Patent application No. Demande de brevet n°

02102829.5

PRIORITY DOCUMENT
SUBMITTED OR TRANSMITTED IN
COMPLIANCE WITH
RULE 17.1(a) OR (b)

Der Präsident des Europäischen Patentamts;
Im Auftrag

For the President of the European Patent Office

Le Président de l'Office européen des brevets
p.o.

R C van Dijk



Anmeldung Nr:
Application no.: 02102829.5
Demande no:

Anmeldetag:
Date of filing: 19.12.02
Date de dépôt:

Anmelder/Applicant(s)/Demandeur(s):

Philips Corporate Intellectual Property GmbH
Habsburgerallee 11
52064 Aachen
ALLEMAGNE
Koninklijke Philips Electronics N.V.
Groenewoudseweg 1
5621 BA Eindhoven
PAYS-BAS

Bezeichnung der Erfindung/Title of the invention/Titre de l'invention:
(Falls die Bezeichnung der Erfindung nicht angegeben ist, siehe Beschreibung.
If no title is shown please refer to the description.
Si aucun titre n'est indiqué se référer à la description.)

System und Verfahren zum Auslesen von Daten

In Anspruch genommene Priorität(en) / Priority(ies) claimed / Priorité(s)
revendiquée(s)
Staat/Tag/Aktenzeichen/State/Date/File no./Pays/Date/Numéro de dépôt:

Internationale Patentklassifikation/International Patent Classification/
Classification internationale des brevets:

G06F17/30

Am Anmeldetag-benannte Vertragsstaaten/Contracting states designated at date of
filing/Etats contractants désignées lors du dépôt:

AT ~~BE~~ BG CH CY CZ DE DK EE ES FI FR GB GR IE IT LI LU MC NL
PT SE SI SK TR

~~Die o.g. Angaben~~ wurden wie folgt geändert :

~~Anmelder~~ für DE : Philips Intellectual Property & Standards GmbH., Steindamm 94,
20099 Hamburg, Deutschland.

Anmelder für alle andere zur Zeit der Einreichung der Anmeldung benannte Vertrags-
staaten : Koninklijke Philips Electronics N.V., Groenewoudseweg 1, 5621 BA Eindhoven,
~~Niederlande~~.

Die ~~Eintragung~~ der geänderten Daten ist mit Wirkung vom 13.03.2003 erfolgt.

BESCHREIBUNG

Verfahren und System zum Auslesen von Daten

Die Erfindung betrifft ein System und ein Verfahren zum Auslesen von Daten aus einem Dokument und ein Verfahren zum Erfassen von Daten.

- 5 Eine Vielzahl von Daten sind heute in elektronischer Form gespeichert. Dies umfasst einerseits die Speicherung in Datenbanken zur gezielten Abfrage eines oder mehrerer Datensätze mittels eines Computers. Andererseits umfasst dies aber auch Dokumente, die zum Abruf und zur Betrachtung durch menschliche Benutzer vorgesehen sind, wie
- 10 HTML- oder XML-Dateien, Tabellen, strukturierte Texte oder Arbeitsblätter einer Tabellenkalkulation. Wie die vorgenannten Datenbanken sind die letztgenannten elektronischen Dokumente zwar auch computerlesbar. Für die Abfrage einzelner Daten aus einem solchen Dokument (bspw. das Auslesen eines einzelnen Eintrags aus einer Tabelle) gibt es jedoch kein spezielles Abfrage-Interface. Das automatische Auslesen von
- 15 Daten aus solchen Dokumenten, die hier – im Gegensatz zu Datenbanken – als schwach strukturiert bezeichnet werden sollen, erfordert üblicherweise die Erstellung einer Abfrage-Anweisung in einer geeigneten Computersprache, bspw. ein PERL-Skript oder einen regulären Ausdruck, der von bekannten Programmen wie grep, sed oder awk interpretiert wird. Die Erstellung einer derartigen Anweisung erfordert Programmier-
- 20 kenntnisse und ist für den Benutzer wenig komfortabel.

- Das Ziel des Auslesens von Daten aus einem Dokument ist in der Regel nicht lediglich ein einmaliges Auslesen der Daten. Sondern es werden bevorzugt mehrfach mit zeitlichem Abstand Daten aus solchen Dokumenten ausgelesen, die sich mit der Zeit ändern,
- 25 d. h. aktualisiert werden. Bspw. kann ein mit einer bestimmten Adresse (URL) in einem Computer-Netzwerk bezeichnetes Dokument (z. B. eine HTML-Seite) eine stets aktuelle Tabelle mit Wetterdaten verschiedener Städte aufweisen. Hierbei werden die angezeigten Daten, bspw. die Temperatur, sich von Tag zu Tag unterscheiden. Unter Umständen ist es sogar möglich, dass sich die absolute Lage eines Bereichs durch Umfor-

matierung ändert. So kann bspw. der Temperaturwert von Paris an einem Tag in der dritten Zeile in Spalte 2 stehen, und an einem anderen Tag in der zweiten Spalte einer anderen Zeile.

- 5 Die hier besonders interessierende Klasse von Dokumenten, bei denen die darin enthaltenen Informationen sich mit der Zeit ändern können, soll hier als "volatil" bezeichnet werden.

Es ist Aufgabe der Erfindung, ein System und ein Verfahren zum Auslesen von Daten aus einem Dokument und ein Verfahren zum Erfassen von Daten anzugeben, mit dem ein Benutzer auf einfache Weise eine Abfrage-Anweisung erstellen kann, mit der Daten, insbesondere aus schwach strukturierten, volatilen Dokumenten abgefragt werden können.

15 Diese Aufgabe wird gelöst durch ein System nach Anspruch 1, ein Verfahren zum Auslesen von Daten nach Anspruch 9 und ein Verfahren zum Erfassen von Daten nach Anspruch 10. Abhängige Ansprüche beziehen sich auf vorteilhafte Ausführungsformen der Erfindung.

20 Bei der erfindungsgemäßen Lösung legt der Benutzer durch eine von einem Computer mit einem darauf ablaufenden Programm verarbeitete Eingabe für mindestens ein Dokument fest, in welchem Bereich sich die ihn interessierenden Daten befinden. Mit Hilfe des Programms wird automatisch eine Abfrage-Anweisung hierfür generiert.

25 Der Computer hat Zugriff auf mindestens ein Dokument. Bevorzugt ist der Computer an ein Computer-Netzwerk, bspw. das Internet angeschlossen, und greift über das Computer-Netzwerk auf ein entfernt liegendes Dokument zu.

Das auf dem Computer ablaufende Programm zeigt bspw. das Dokument an und fordert den Benutzer auf, den interessierenden Datenbereich festzulegen, bspw. durch Markierung mit einer Zeigereinheit (wie einer Maus). Optional kann der Benutzer zusätzlich auch eine zweite Eingabe vornehmen, mit der ein oder mehrere weitere Bereiche, hier

bezeichnet als Strukturbereiche, festgelegt werden, die beim Auffinden des gewünschten Bereichs hilfreich sein können. Bspw. kann es sich bei diesen Strukturbereichen um Zeilen- oder Spaltenköpfe handeln, die in einer Tabelle zu einer Zelle mit dem gewünschten Inhalt führen.

- 5 Erfindungsgemäß wird anhand der Eingaben des Benutzers automatisch eine Abfrage-Anweisung erzeugt. Die Abfrage-Anweisung wird in einer für einen Computer lesbaren und durch ein geeignetes Interpreter-Programm ausführbaren Form geliefert und bevorzugt abgespeichert. Beim Ausführen der Abfrage-Anweisung wird der festgelegte Datenbereich des Dokuments ausgelesen. Ist vom Benutzer zusätzlich ein Strukturbereich vorgegeben, so enthält die erstellte Abfrage-Anweisung bevorzugt eine Adressierungsanweisung entsprechend der Position oder des Inhalts des vorgegebenen Strukturbereichs.
- 10
- 15 Gemäß einer Weiterbildung der Erfindung ist für die Abfrage-Anweisung eine spezielle Grammatik vorgegeben. Hierbei setzt sich ein gültiger Ausdruck aus einer vorgegebenen Reihenfolge von Terminalzeichen zusammen. Die Grammatik enthält bevorzugt Adressierungsanweisungen, mit denen bestimmte Positionen eines Dokuments einerseits absolut (bspw.: Dokumentanfang) und andererseits auch relativ zu einem vorher
- 20 bezeichneten Bereich (bspw.: zwei Zeilen tiefer) adressiert werden.

- Die verwendete Grammatik ist bevorzugt von einfacher Struktur. Hierbei ist die Grammatik bevorzugt auf den Typ des interessierenden Dokuments abgestimmt. Bspw. kann eine spezielle Grammatik zur Adressierung in Fließtexten vorgesehen sein, die dann
- 25 bspw. eine Adressierung auf Wort- und Satzbasis vorsehen kann (bspw.: zweites Wort im dritten Satz). Oder es kann eine spezielle Grammatik für Tabellen vorgesehen sein, mit der eine Adressierung auf Zeilen- und Spaltenbasis besonders gut möglich ist (bspw.: drittes Feld in der Zeile, die mit "Paris" beginnt).

- 30 Gemäß einer Weiterbildung der Erfindung erfolgt die automatische Erstellung der Abfrage-Anweisung, indem mehrere gültige Abfrage-Anweisungen der vorgegebenen Grammatik erzeugt werden und diese Abfrage-Anweisungen daraufhin überprüft wer-

den, ob bei ihrer Ausführung der interessierende Datenbereich des Dokuments ausgelesen wird. Eine der erfolgreichen Abfrage-Anweisungen wird ausgewählt, bspw. anhand eines Komplexitätskriteriums.

- 5 Bevorzugt erfolgt die Erstellung einer automatischen Abfrage-Anweisung nicht lediglich anhand eines Dokumentes, sondern es werden mehrere Trainingsdokumente verarbeitet. So kann die Wahrscheinlichkeit erhöht werden, dass die automatisch erstellte Abfrage-Anweisung auch bei volatilen Dokumenten stets die gewünschten Daten liefert, ohne dass eine ständige Anpassung erforderlich ist.

10

Eine einmal erstellte und bevorzugt mit mehreren Trainingsdokumenten verifizierte Abfrage-Anweisung wird bevorzugt abgespeichert. Sie kann dann im zeitlichen Abstand wiederholt ausgeführt werden, um aus dem adressierten Bereich eines ständig aktualisierten Dokuments stets den aktuellen Wert auszulesen. Der Wert kann auf viele

- 15 verschiedene Arten weiterverarbeitet werden. So können bspw. aktuelle Informationen von verschiedenen, über ein Computer-Netzwerk abrufenbaren Dokumenten gesammelt und zu einer nach persönlichen Präferenzen zusammengestellten Informationssammlung verarbeitet werden.

- 20 Nachfolgend wird ein Ausführungsbeispiel der Erfindung anhand von einer Zeichnung näher beschrieben. Hierbei zeigt:

Fig. 1 einen Graph zur Herleitung einer Abfrage-Anweisung.

- 25 In dem Ausführungsbeispiel sollen in ein automatisch erstelltes persönliches Radioprogramm, wie es bspw. in der WO 99/39466 beschrieben ist, aktuelle Wetterinformationen eingefügt werden. Die hierfür benötigten Wetter-Informationen sind auf verschiedenen Internet-Seiten (HTML-Dokumenten) ständig aktuell verfügbar. Dem Benutzer soll die Möglichkeit gegeben werden, mittels eines Computers auf einfache Weise die
- 30 Informationen festzulegen, die dann in seinem täglichen persönlichen Radioprogramm (mittels Sprachsynthese) eingeblendet werden sollen.

- Hierfür wird ein Computersystem mit Eingabemitteln (bspw. Tastatur, Maus) und Ausgabemitteln (bspw. Monitor) verwendet. Der Computer ist an das Internet angeschlossen. Auf dem Computer ist ein Programm installiert, mit dem der Benutzer auf einfache
- 5 Weise eine Abfrage-Anweisung für die ihn interessierenden Daten formulieren und diese an den Dienstleister übermitteln kann, der die Zusammenstellung des persönlichen Radioprogramms für ihn durchführt. Die Funktion dieses Programms wird nachfolgend im Detail beschrieben.
- 10 Für die zu formulierende Abfrage ist eine Grammatik vorgegeben. Diese kann beliebig definiert werden. Eine solche Grammatik umfasst Terminalzeichen der folgenden Arten:
1. Absolute Adressierung, um einen absolut festgelegten Bereich innerhalb eines Do-
15 kuments zu adressieren (bspw. TOP, BOTTOM, ROOT).
 2. Relative Adressierung, um Orte oder Bereiche innerhalb eines Dokuments ausgehend von einem Ursprungsort oder -bereich zu adressieren (bspw. next_paragraph, previews_word, next_list_item, cell_up, to_first_row, parent_node, first_child,
20 next_sibling).
 3. Suchkommandos, um Orte bedingt zu adressieren. Ein Suchkommando besteht aus einem Suchbereich (bspw. in_paragraph, in_subtree, within_column) und einer Bedingung (bspw. contains_text (T), has_format (F), is_a_number, is_smaller_than
25 (n), carries_xml_tag (T)), ggfs. einen relativen Pfad zu dem Ort, wo die Bedingung zutreffen soll (bspw. die relative Adressierung wie oben unter 2.) und einen Indikator, mit dem bei mehreren Treffern ein Einzelner ausgewählt werden kann (bspw. first_occurrence, last_occurrence, nth_occurrence (n)).
- 30 ~~Man~~ Eine Kombination von mehreren der oben genannten Terminalzeichen lassen sich Abfrage-Anweisungen für verschiedene Dokumente aufbauen. Bspw. ist für den Fachmann ersichtlich, dass eine Abfrage-Anweisung der folgenden Art aufgebaut werden

kann: "Nimm die dritte Tabelle des Dokuments und wähle dort das erste Vorkommen einer Zahl, die von einem \$-Zeichen gefolgt ist in einer Zelle, die sich in der Reihe befindet, deren erster Spalteneintrag "Kanada" lautet". Eine solche Abfrage-Anweisung könnte beispielsweise wie folgt lauten:

5

```
TOP to_next_table to_next_table to_next_table
find (in_table, is_a_number and has_format ("$$")
      and (to_first_column contains_text ("Kanada")),
      first_occurrence)
```

10

Das auf dem Computer ablaufende Programm nimmt Eingaben des Benutzers entgegen, mit denen in einem Dokument die interessierenden Daten bezeichnet werden. Das Programm erstellt dann automatisch eine Abfrage-Anweisung. Die Abfrage-Anweisung ist in der jeweils vorgegebenen Grammatik formuliert. Bei entsprechender Ausführung, d.h. Durchlauf durch einen entsprechenden Interpreter, in dem jeweiligen Dokument liefert sie die bezeichneten Daten.

Bspw. wohnt der Benutzer in Frankfurt und möchte im Rahmen seines persönlichen Radioprogramms jeden Tag über die dortigen aktuellen meteorologischen Werte wie Temperatur und Luftfeuchtigkeit informiert werden. Er sucht eine im Internet abrufbare HTML-Seite auf, in der diese Informationen täglich aktuell angegeben werden. Die nachfolgende Tabelle gibt beispielhaft den Inhalt einer solchen Seite wieder:

25

Ort	Temperatur (°C)	Feuchtigkeit (%)	Wolken (%)
Aachen	24	90	80
Berlin	18	70	30
<u>Frankfurt</u> a.M.	22	<u>60</u>	20
Köln	23	50	95

Um die Informationen über die aktuelle Feuchtigkeit in das persönliche Radioprogramm einzufügen, erstellt der Benutzer an dem Computer mit dem Programm eine Abfrage,

die dem Dienstleister zur Erstellung des persönlichen Radioprogramms übermittelt wird. Bei der Ausführung des Programms ruft der Benutzer das Dokument mit der oben angegebenen Tabelle ab. Er markiert den interessierenden Wert, hier den Wert für die relative Luftfeuchtigkeit in Frankfurt (60, die Markierung ist als Unterstreichung dargestellt) mit der Maus. Zusätzlich markiert der Benutzer den Zeilenkopf ("Frankfurt") als Strukturbereich, der bei der Adressierung des interessierenden Werts verwendet werden kann.

Aus diesen Informationen erstellt das Programm automatisch eine Abfrage-Anweisung nach der jeweils vorgegebenen Grammatik. Nachfolgend ist die Arbeitsweise des Programms als Pseudo-Code angegeben:

1. SET TargetExpression := <empty>
 SET DocumentsAndMarksList := <empty>
- 15 2. FOR d IN { Alle Trainingsdokumente} DO
3. IF TargetExpression eine Abfrageanweisung ist , die zu deinem gültigen Eintrag
 im Trainingsdokument d führt
20 THEN Bereich anzeigen, auf den TargetExpression zeigt
 Benutzer fragen, ob der angezeigte Bereich den gewünschten Daten entspricht
 IF Benutzer antwortet mit "ja" THEN GOTO 5.
4. Den Benutzer um Eingabe bitten, mit der der gewünschte Bereich im
25 Trainingsdokument d markiert wird. Optional kann der Benutzer zusätzlich eine
 oder mehrere weitere Markierungen von Strukturbereichen vornehmen, die bei der
 Abfrage berücksichtigt werden sollen (Wenn er dies nicht tut, SET A :=
 <empty>)
- 30 5. Das Tripel (d, M, A) an DocumentsAndMarksList anhängen
6. FOR alle Abfrageanweisungen L, die aus der Grammatik G herleitbar sind und
 die eine vorgegebenes Komplexitätsmaß nicht überschreiten

7. SET count := 0
8. FOR alle Tripel (t.A, t.M, t.D) IN DocumentsAndMarksList DO
- 5 9. IF (MARKING_DUE_TO_LOCATOR_EXPR(t.D, L) = t.M)
AND (t.A \subseteq LOCATOR_EXPRESSION_PATH (t.D, L))
THEN count++
- 10 10. DONE (Nächstes Tripel in Schritt 8 bearbeiten)
11. IF (count > bestcount)
OR ((count == bestcount) AND (COMPLEXITY(L) < COMPLEXITY(bestL))
THEN
- 15 SET bestL := L ; SET bestcount := count
12. DONE (mit der nächsten Abfrageanweisung in Schritt 6 fortfahren)
13. DONE (mit dem nächsten Dokument in Schritt 2 fortfahren)
- 20 14. RETURN bestL

Dieses Programm benutzt die folgenden Funktionen:

- 25 MARKING_DUE_TO_LOCATOR-EXPR(Dokument d, Abfrageanweisung):
Diese Funktion interpretiert die Abfrageanweisung und gibt die Daten des Dokuments d zurück, die an der durch die Abfrageanweisung bezeichneten Stelle stehen.

- LOCATOR_EXPRESSION_PATH (Dokument d, Abfrageanweisung):
30 Diese Funktion gibt einen Satz von Bereichen zurück, die durchlaufen werden, wenn die Abfrageanweisung im Dokument d ausgeführt wird.

COMPLEXITY(Abfrageanweisung L):

Ein Komplexitätsmaß für die Abfrageanweisung L, beispielsweise die Länge des

Ausdrucks. Dieses Komplexitätsmaß wird zur Auswahl verwendet, wenn sonst mehrere Abfrageanweisungen gleichwertig sind.

Das Programm arbeitet mit einer Anzahl von Trainingsdokumenten. Hierbei handelt es
5 sich bevorzugt um verschiedene Dokumente, die unter derselben URL zu verschiedenen
Zeitpunkten abrufbar waren, so dass durch den zur Verfügung stehenden Satz an Trainingsdokumenten möglichst gut die Volatilität des interessierenden Dokuments abgedeckt wird. Das Programm kann aber auch benutzt werden, wenn nur wenige Trainingsdokumente, oder sogar nur ein einzelnes Trainingsdokument zur Verfügung stehen. Die
10 äußere Schleife (2. – 13.) wird dann nur entsprechend wenige Male durchlaufen.

Das oben als Pseudo-Code dargestellte Programm arbeitet wie folgt:

Eine äußere Schleife (2. – 13.) wird für alle vorliegenden Trainingsdokumente durch-
15 laufen. Liegt bereits eine Abfrage-Anweisung TargetExpression vor, die auf einen gültigen Eintrag führt, so wird dieser Bereich markiert und der Benutzer gefragt, ob dies der gewünschte Bereich ist (Schritt 3). Sonst wird der Benutzer gebeten, seinerseits den gewünschten Bereich (und optional zusätzlich einen oder mehrere Strukturbereiche) zu markieren (Schritt 4).

20

Die Tripel aus Trainingsdokumenten, gewünschten Bereichen und (optional) Strukturbereichen werden an eine Liste DocumentsAndMarksList angehängt (Schritt 5).

Im Schritt 6 werden aus der Grammatik G eine Anzahl von Abfrage-Anweisungen L
25 erzeugt. Bevorzugt sind dies sämtliche gültigen Ausdrücke der Grammatik G, die ein vorgegebenes Komplexitätsmaß (bspw. die Gesamtlänge des Ausdrucks) nicht überschreiten. Aus der Definition der Grammatik ist es für den Fachmann leicht möglich, automatisch gültige Ausdrücke zu erzeugen.

30 Für jede der erzeugten Abfrage-Anweisungen L werden alle vorliegenden Tripel (Dokumente mit Zielbereichen und ggf. Strukturbereichen) überprüft, ob der Ausdruck zum

gewünschten Ergebnis führt. Ist dies der Fall, wird ein Zähler (count) hochgezählt (Schritte 8 – 10).

Der Erfolg einer Abfrage-Anweisung, d. h. die Anzahl korrekter Markierungen in den
 5 zur Verfügung stehenden Trainingsdokumenten, wird mit dem bisher höchsten Erfolg (Best Count) verglichen. Bei einem höheren Wert wird der aktuelle Ausdruck als bester Kandidat behalten. Bei einem Gleichstand mit dem bisherig besten Kandidaten wird der Ausdruck mit der niedrigsten Komplexität behalten (Schritt 11).

10 Am Ende wird der auf diese Weise ermittelte beste Ausdruck als gefundene Abfrage-Anweisung zurückgegeben (Schritt 14).

Nachfolgend soll der Ablauf des Programms anhand eines einfachen Beispiels illustriert werden:

15

Nachfolgend ist beispielhaft einfache Grammatik angegeben, die beispielsweise für Tabellen-Strukturen verwendet werden könnte. Terminalzeichen sind in Kleinschreibweise, Nicht-Terminalzeichen in Großbuchstaben angegeben:

20 `EXPRESSION ::= top_left_cell ROW_HEADER_SEARCH_EXPR`

`ROW_HEADER_SEARCH_EXP ::= find (within_column, contains_text(#))
 ROW_ELEMENT_SELECTION`

25 `ROW_ELEMENT_SELECTION ::= select_entire_cell |
 cell_left ROW_ELEMENT_SELECTION`

Wird der oben angegebene Algorithmus mit dieser Grammatik auf die oben dargestellte Tabelle angewendet, in der als Zielbereich die Zahl 60 und als Strukturbereich das Wort
 30 "Frankfurt" markiert sind, so könnte bspw. die gelieferte Abfrage-Anweisung

TOP find (within_column, contains_text ("Frankfurt"))
cell_right cell_right select_entire_cell

- 5 wie in Fig. 1 dargestellt aus der Grammatik abgeleitet werden. Hierbei wird der Inhalt "Frankfurt" des markierten Strukturbereichs in eine "find"-Anweisung umgesetzt, mit der in der ersten Spalte nach dem Wort "Frankfurt" gesucht wird (das Zeichen "#" in der Grammatik wird durch den Inhalt des selektierten Strukturbereichs "Frankfurt" ersetzt).

- 10 Ausgehend von der gefundenen Zelle befindet sich die gesuchte Zelle mit dem Inhalt "60" zwei Zellen weiter rechts, d. h. sie wird durch zwei Aufrufe von cell_right erreicht. Die so gefundene Zelle wird als ganzes markiert und liefert den gewünschten Inhalt "60".

PATENTANSPRÜCHE

1. System zum Auslesen von Daten aus einem Dokument mit
 - mindestens einem Computer mit Anzeige- und Eingabemitteln zur Bedienung durch einen Benutzer,
 - und Mitteln zum Zugriff auf mindestens ein Dokument,
 - 5 - und Programm-Mitteln, durch die mindestens eine erste Eingabe des Benutzers verarbeitet wird, mit der ein auszulesender Datenbereich des Dokuments festgelegt wird,
 - wobei durch die Programm-Mittel automatisch eine Abfrage-Anweisung zum Auslesen des festgelegten Datenbereiches erstellt wird, so dass bei Ausführung
 - 10 der Abfrage-Anweisung der festgelegte Datenbereich des Dokuments ausgelesen wird.
2. System nach Anspruch 1, bei dem
 - die Programm-Mittel auch eine zweite Eingabe des Benutzers verarbeiten, mit
 - 15 der mindestens ein Strukturbereich des Dokuments festgelegt wird, der bei der Festlegung des Datenbereichs verwendet wird,
 - wobei die von den Programm-Mitteln erstellte Abfrage-Anweisung eine Adressierungsanweisung entsprechend der zweiten Eingabe enthält.
- 22 3. System nach einem der vorangehenden Ansprüche, bei dem
 - für die Abfrage-Anweisung eine Grammatik vorgegeben ist,
 - wobei die Grammatik mindestens eine absolute Adressierungsanweisung und mindestens eine relative Adressierungsanweisung aufweist.

4. System nach einem der vorangehenden Ansprüche, bei dem
- bei der Erstellung der Abfrage-Anweisung eine Anzahl von gültigen Abfrage-Anweisungen einer vorgegebenen Grammatik erzeugt werden,
- 5
- und diese Abfrage-Anweisung daraufhin überprüft werden, ob bei ihrer Ausführung der gewünschte Datenbereich des Dokuments ausgelesen wird,
 - wobei unter den Abfrage-Anweisungen, bei deren Ausführung der gewünschte Datenbereich des Dokuments ausgelesen wird, eine Abfrage-Anweisung ausgewählt wird.
- 10
5. System nach Anspruch 4, bei dem
- die Abfrage-Anweisung nach einem Komplexitätskriterium ausgewählt wird.
6. System nach einem der vorangehenden Ansprüche, bei dem
- 15
- der Computer auf mehrere Dokumente zugreift,
 - wobei eine oder mehrere Abfrage-Anweisungen überprüft werden, bei wie vielen der Dokumente der bei ihrer Ausführung ausgelesene Bereich die vom Benutzer gewünschten Daten enthält.
- 20
7. System nach einem der vorangehenden Ansprüche, bei dem
- der Computer an ein Computer-Netzwerk angeschlossen ist, so dass er über das Computer-Netzwerk auf Dokumente zugreifen kann.
8. System nach einem der vorangehenden Ansprüche, bei dem
- 25
- Mittel zum Abspeichern der Abfrage-Anweisung vorhanden sind.

9. Verfahren zum Auslesen von Daten aus einem Dokument, bei dem

- auf mindestens einem Computer mit Anzeige- und Eingabemitteln zur Bedienung durch einen Benutzer und Mitteln zum Zugriff auf ein Dokument,
- Programm-Mittel ausgeführt werden, durch die mindestens eine erste Eingabe des Benutzers verarbeitet wird, mit der ein auszulesender Datenbereich des Dokuments festgelegt wird,
- wobei durch die Programm-Mittel automatisch eine Abfrage-Anweisung zum Auslesen des festgelegten Datenbereiches erstellt wird, so dass bei Ausführung der Abfrage-Anweisung aus dem Dokument der festgelegte Datenbereich ausgelesen wird.

10. Verfahren zum Erfassen von Daten, bei dem

- mit dem Verfahren gemäß Anspruch 9 eine Abfrage-Anweisung erstellt wird,
- und die Abfrage-Anweisung abgespeichert wird,
- und die Abfrage-Anweisung im zeitlichen Abstand wiederholt ausgeführt wird.

ZUSAMMENFASSUNG

System und Verfahren zum Auslesen von Daten

Beschrieben werden ein System und ein Verfahren zum Auslesen von Daten aus einem Dokument und ein Verfahren zum Erfassen von Daten. Um Daten, die in volatilen (d. h. mit der Zeit veränderlichen) und schwach strukturierten Dokumenten (bspw. HTML-Seiten) vorhanden sind auszulesen, wird ein System mit einem Computer mit Anzeige- und Eingabemitteln zur Bedienung durch einen Benutzer vorgeschlagen. Der Computer kann auf mindestens ein Dokument zugreifen. Auf dem Computer läuft ein Programm ab, durch das mindestens eine erste Eingabe des Benutzers verarbeitet wird, mit der ein auszulesender Datenbereich des Dokuments festgelegt wird. Das Programm erstellt automatisch eine Abfrage-Anweisung zum Auslesen des festgelegten Datenbereichs. Die Abfrage-Anweisung kann bspw. gespeichert werden. Bei Ausführung der Abfrage-Anweisung wird der festgelegte Datenbereich des Dokuments ausgelesen. Dies kann im zeitlichen Abstand wiederholt durchgeführt werden.

Fig. 1

Fig. 1

